U.S. DEPARTMENT OF COMMERCE
National Technical Information Service

PB-252 745-T

APPLICATION OF SOME PATTERN RECOGNITION METHODS IN INTERPRETING MASS SPECTRA

Environmental Protection Agency

1974

PB 252 745-T

TR 490-74

141030

K. Varmuza

APPLICATION OF SOME PATTERN RECOGNITION METHODS IN INTERPRETING MASS SPECTRA

[Anwendung einiger Methoden der automatischen Zeichenerkennung auf die Interpretation von Massenspektren]

Monatshefte für Chemie, Vol. 105, 1974, pp. 1-10

Translated for the
ENVIRONMENTAL PROECTION AGENCY
by
Translation Consultants, Ltd., Arlington, Virginia

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U. S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

BIBLIOGRAPHIC DATA SHEET	1. Report No. EPA TR 490-74	2.	3. Recipient's Accession No.
4. Title and Subtitle Application of some pattern recognition methods interpreting mass spectra			5. Report Date 1974 6.
7. Author(s) K. Varmuza			8. Performing Organization Rept. No.
9. Performing Organization Name and Address Translated fromthe German by EPA			10. Project/Task/Work Unit No. 11. Contract/Grant No.
12. Sponsoring Organization Name and Address EPA ISD Translation Services Division Library RTP, NC 27711			13. Type of Report & Period Covered 14.

15. Supplementary Notes

Monatshefte fur Chemie 105:1-10 1974

16. Abstracts

17. Key Words and Document Analysis. 17a. Descriptors

Chemical Analysis
Computers

17b. Identifiers/Open-Ended Terms

Chemical Test
Computer software

17c. COSATI Field/Group 09-B

18. Availability Statement	19. Security Class (This Report)	21. No. of Pages
Exclusive permission granted to NTIS for the storage and distribution of this material	UNCLASSIFIED	15
	20. Security Class (This Page) UNCLASSIFIED	22. Price 3.50

K. Varmuza*

APPLICATION OF SOME PATTERN RECOGNITION METHODS IN INTERPRETING MASS SPECTRA

Introduction

Automated and computerized large-capacity equipment, which is used in instrumental analysis, produce a high amount of data during an analysis. This flood of data often contains much valuable information which is hard to interpret. The most important task of the computer therefore is not the production of data but data reduction, and a useful representation of the results. The great amount of data (e.g. spectra) which accumulates in the course of time, can hardly be surveyed and taken into consideration in every analysis even by the expert. Therefore it is reasonable that the chemical-physical interpretation of data, at least an initial interpretation, is performed by a computer program. However, such interpretation encounters considerable difficulties. The information desired (e.g. the chemical structure of an unknown substance) usually is contained in the available data in a coded form only (e.g. spectra). The theoretical relationship between the required information and the present structure of the data generally is not as well known as is necessary for a mathematical formulation. Programs were developed which derive a chemical structure from the mass spectrum and the NMR-spectrum [1, 2] by applying strategies which were tested in chess game computer programs and which came close to a chemist's working method in the spectrum evaluation. These heuristic methods are very efficient, but they require very extensive and time-consuming computer programs, which hardly will be available to a greater number of users in the near future.

---

* Reported in abridged form at a lecture session of the Society of Austrian Chemists and of the Society of the Chemical Industry, 26 to 29 September 1973, Linz.

The other way to interpret the data is purely empirical, without consideration of the physical-chemical relationships. For that purpose, methods can be used which were developed in the area of automatic recognition, for character recognition or for photographic interpretation, for example. The basic method of operation is always the following. A classifier (this is basically an arithmetic operation) is derived from data collection using mathematical methods. This classifier correctly classifies available data with the smallest error rate possible [3, 4]. (A classifier, that can process mass spectra in a large spectrum library and decide if an unknown substance has a benzene ring or not is a case in point). If the information required is contained in the data and a useful classification system has been selected, it can be assumed that the classifier will correctly classify new data with a high percentage of accuracy. Many methods of automatic character recognition are very simple and clear, and only require a basic knowledge of mathematics. The determination of a classifier generally is connected with extensive, but uncomplicated mathematical operations, but the application of the classifier can be accomplished simply and quickly.

This paper will discuss some methods of automatic recognition for those interested, in a manner that will make it easy to understand the geometric considerations. These methods were used, insofar as they were suited to automatic interpretation of low resolution mass spectra. The purpose was to develop classifiers, capable of providing direct references to the chemical structure of a low molecular substance from a low resolution mass spectrum.

Representation of Characteristic Amounts as a Vector

We start with the fact that a series of characteristics (measurement results) are present from a large number of known objects (substances). Characteristics x_1, x_2, ... x_n of substance x are the physical-chemical data for example, or in the case of a mass spectrum, the peak heights in the spectrum with mass numbers 1 to n. The characteristics of a substance are combined in characteristics vector $\vec{X}$. An n-dimensional space is needed to represent this vector. The vector is drawn from its origin and the head of the vector has coordinates x_1, x_2, ... x_n. In what follows only the head of the vector will be taken into consideration. Generally it is enough to picture a two- or three-dimensional similar case in order to understand the classification method.

Every low resolution mass spectrum can be described as a point in an n-dimensional space (spectrum space) without any loss of information. Each coordinate axis corresponds to a mass number. The coordinates of the head correspond to the peak heights in the spectrum.

The working hypothesis for the development of a classifier is as follows. It is hoped that the mass spectra of chemically similar substances form clusters in this n-dimensional space and that the Euclidian distance beteween the mass spectra is a measure for the chemical similarity of the substances [5]. Euclidian distance d between points X (x_1, x_2, ... x_n) and Y (y_1, y_2, ... y_n) is given by

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots (x_n - y_n)^2}$$

as an analogy for a clearly two- or three-dimensional case. In the following

it will be shown that several methods of automatic pattern recognition can be used, depending on the type of cluster formation.

Classification by Distance Measurement to Prototypes

Figure 1 shows the theoretically possible cluster formation for a two-dimensional case. The rings correspond to class A substance spectra, substances with benzene rings for example; the triangles correspond to class B substance spectra, substances without benzene ring. S_A is the prototype of this class as the center point of all class A spectra. In this simple case it can be decided to which class the spectrum belongs when the distance is measured. It is assigned to class B if the distance of a spectrum to the center point, S_A, is greater than critical distance d_k, otherwise to class A.

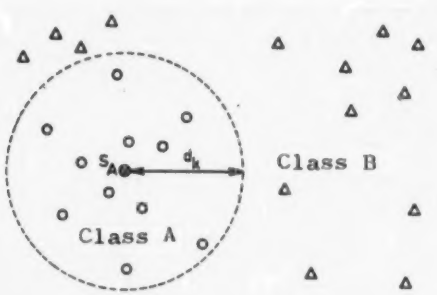

Figure 1. Classification by distance measurement to center point S_A of class A. Only class A members fall within the critical Euclidian distance d_k.

Figure 2 shows the actual relationships for 500 mass spectra. The center point of mass spectra of 113 substances with benzene ring, the center point of mass spectra of 387 substances without benzene ring and the Euclidian

distances d of all 500 spectra to the two center points were calculated. The frequency distributions of the distances show that there is no critical distance which adequately separates the two classes. Normalization of the peak heights in the mass spectra to the same sum of peak heights in each spectrum, or normalization to the same vector length (the heads of the vector then are on a hypersphere [6]) did not make better separation of the two classes possible. On the other hand, using logarithmic peak heights instead of the customary linear values produced good results. This simple method of distance determination for mass spectra seems to be of little value because the classification of other chemical structures (aliphatic alcohol, oxygen in the molecule) brought similar results. It can be quite successful, however, in other data structures with a more fully developed cluster formation. Also weight [3] of the individual characteristics to help in differentiating the classes should be taken into consideration.

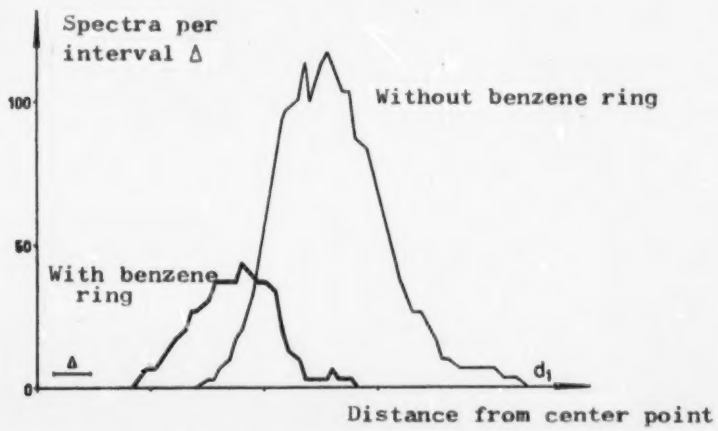

Figure 2a.

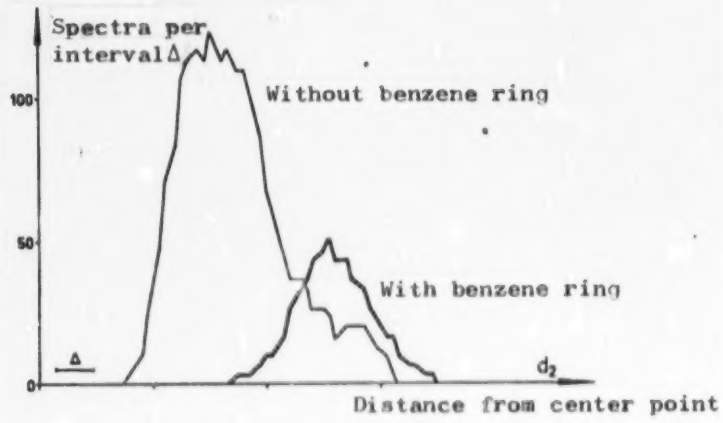

Figure 2b.

Figure 2. Frequency distribution of the Euclidian distances d_1 and d_2 of 500 mass spectra to center points of substances with and without benzene ring. d_1 is the distance of center point of substances with benzene rings, d_2 is the distance of the center point of the substances without benzene ring. (The measure is linear in arbitrary units). The logarithms of the mass spectra peak heights were taken and were normalized to the base peak.

Classification by Decision Planes

Figure 3 shows a cluster formation in which the center point method is not used with success. However, it is possible to find a straight line (in the multidimensional case, a hyper plane) which separates the two classes. This method of automatic pattern recognition was the first one tested with mass spectra [7, 8, 9]. The required decision planes are found, if they exist, by an iteration method (designated as "learning machine" in American publications), which several times has been described in detail [7, 10, 11]. The classifier contains the position of the plane in form of a decision vector; a spectrum is assigned to class A or B, depending on which side of the decision plane it is.

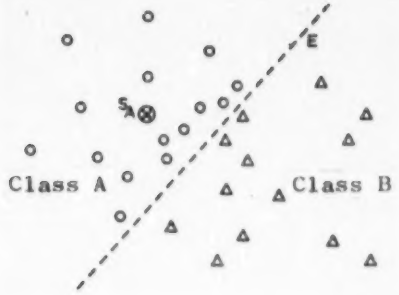

Figure 3. Classification by decision plane E, which completely separates classes A and B. Center point, S_A, class A is not suited for classification by distance measurement.

Table 1 contains several own results, which were obtained by this classification method. Decision planes were calculated with the help of 250 mass spectra (the so-called training set), which completely separate two classes (alkyl benzenes from all other compounds). The usefulness of the decision planes was checked with 250 other mass spectra (the so-called prediction set). Each classifier provides the answer "yes" or "no" for a specific chemical structure. Reliability of the answers presented in Table 1 shows the percentage of correct answers in the classification of the prediction set. Extensive investigations [12] show that only few structural characteristics can be used for this classification method, as compared to earlier optimistic statements [9, 10]. The six classes presented in Table 1 from among 31 chemical structures investigated had a higher than 70% answer reliability.

Table 1. Classification by decision planes

Chemical structure (class)	Percentage of class in prediction set (%)	Reliability of answers in classification of prediction set (%)	
		yes	no
alkylbenzene	26	75	99
benzene ring	22	91	97
>CO on benzene ring	5	71	99
-OH on benzene ring	4	80	99
oxygen in molecule	55	84	83
nitrogen in molecule	24	78	95

Classification Using Neighbor Spectra

Figure 4 shows a less distinctive cluster formation and no simple decision plane can be used to separate the two classes. An attempt could be made to perform a separation by using several straight lines or a function of a higher order, but this is mathematically very complicated. Even so, each point in this distribution can be assumed to be unknown and its class can be determined by a simple method. The nearest neighbor (the so-called first neighbor) indicates the correct class. This logically very simple neighbor method has been known for a long time in the theory of automatic pattern recognition and has been investigated extensively [4, 13, 14].

The neighbor method can be applied in the field of spectrum interpretation when comparing an unknown spectrum with a spectrum library [15]. The criterion of similarity of two spectra is the distance of the vector heads in a multidimensional spectrum space. Initial experiments with mass spectra showed [16], that the Euclidian distance between the spectra actually is in

close relationship with the chemical similarity of the substances in question (this means the smaller the distance, the more structural characteristics are in general agreement). Furthermore it became evident that the first neighbor always has the best classification ability.

Much better results are achieved when several neighbor spectra are used for clasification. The majority of the neighbors decides on the classification of the unknown spectrum. Each spectrum in the library containing 500 mass spectra was assumed to be unknown. The six nearest neighbor spectra of the 449 remaining spectra were used for the classification. The results presented in Table 2 show the light rate of success of this classification method. The "yes" answer of the classifier for the structural characteristic "benzene ring" was correct with 96% and the "no" answer was correct with 99%. In 2% (of 500 spectra) there was a tie of the six neighbors and the classifier reacted with a rejection. It should be mentioned that each mass spectrum was classified by the mass spectra of six other substances, because each substance appears only once in the library of spectra used. In practice, a mass spectrum of the investigated substance very often will be present in the library.

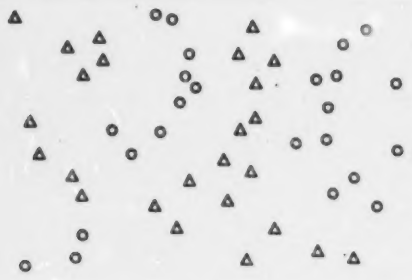

Figure 4. The class of each point can be read from the nearest neighbor in this cluster formation.

Table 2. Classification by six neighbor spectra

Chemical structure (class)	Percentage of class (%)	Reliability of answer (%)		Rejection (%)
		yes	no	
alkane	8	75	99	3
aliphatic alcohol	12	88	93	4
benzene ring	23	96	99	2
pyridine ring	7	77	97	3
>CO	28	86	79	9
$-NR_2$	14	97	92	2
-OH on benzene ring	4	82	98	1
$-CH_3$	78	85	73	9
$-C_2H_5$	40	67	82	19
oxygen in molecule	53	99	77	11
nitrogen in molecule	25	94	90	8

Conclusion

Some of the many methods offered by theorists of automatic pattern

recognition are described in brief. Unfortunately, there is no recognizable connection between data structure and a useful method of classification. Only the practice will show which methods are useful for a certain problem. The methods of automatic pattern recognition (and corresponding computer programs) are so general, that they can be applied to many problems [5]. When programs for different methods of pattern recognition are available then they can be used routinely without great expenditure in all larger data collections. Some of the methods described are suitable in providing direct reference to the chemical structure of compounds from low resolution mass spectra of low molecular compounds. However, automatic interpretation of spectra will not be able to replace the chemist in the clarification of the structure and in the identification of the substance. It supplies additional results which can help to solve some problems after a critical evaluation.

Data and Programs

In this work, 500 low resolution mass spectra were used from spectrum collections [17, 18] which originated from 500 different substances with a molecular weight between 16 and 150 and chemical formulas C_{1-11}, H_{1-22}, O_{0-2} and N_{0-2}. All programs were written in FORTRAN IV. The work was accomplished with an IBM 7040 computer (core storage 32K/36 bit) in the Computer Center of the Vienna Technological Institute.

I would like to thank Prof. Dr. A. Maschka for his assistance. I also would like to thank Dr. P. Krenmayr for valuable discussions and Mr. H. Urban for his active assistance.

REFERENCES

1. Buchanan, B. G., Duffield, A. M., Robertson, A. V., in: Mass Spectrometry-Techniques and Applications (W. A. Milne ed.), New York: Wiley, 1971, p. 121.

2. Crawford, L. R., Morrison, J. D., Anal. Chem., Vol. 43, 1971, p. 1790.

3. Meyer-Brötz, G., Schürmann, J., Methoden der automatischen Zeichenerkennung [Methods of Pattern Recognition], Muenchen, Oldenburg, 1970.

4. Nilsson, N. J., Learning Machines, New York: McGraw-Hill, 1965.

5. Kowalski, B. R., Bender, C. F., J. Amer. Chem. Soc., Vol. 94, 1972, p. 5632.

6. Crawford, L. R., Morrison, J. D., Anal. Chem., Vol. 40, 1968, p. 1469.

7. Jurs, P. C., Kowalski, B. R., Isenhour, T. L., Anal. Chem., Vol. 41, 1969, p. 21.

8. Jurs, P. C., Kowalski, B. R., Isenhour, T. L., Reilley, C. N., Anal. Chem., Vol. 41, 1969, p. 690.

9. Jurs, P. C., Kowalski, B. R., Isenhour, T. L., Reilley, C. N., Anal. Chem., Vol. 42, 1970, p. 1387.

10. Isenhour, T. L., Jurs, P. C., Anal. Chem., Vol. 43, 1971, p. 20 A.

11. Krenmayr, P., Varmuza, K., Allgem. Prakt. Chem., Vol. 23, 1972, p. 289.

12. Varmuza, K., Krenmayr, P., Z. Anal. Chem., Vol. 266, 1973, p. 274.

13. Cover, T. M., Hart, P. E., IEEE Trans., IT 13, 1967, p. 21.

14. Meisel, W. S., Computer-Oriented Approaches to Pattern Recognition, New York-London: Academic Press, 1972.

15. Kowalski, B. R., Bender, C. F., Anal. Chem., Vol. 44, 1972, p. 1405.

16. Varmuza, K., Z. Anal. Chem., 1973, in print.

17. Stenhagen, E., Abrahamsson, S., McLafferty, F. W., Atlas of Mass Spectral Data, New York: Interscience, 1969.

18. Benz, W., Massenspektrometrie organischer Verbindungen [Mass Spectrometry of Organic Compounds], Frankfurt a.M., Akad. Verlagsges, 1969.